CHAPTER FIFTEEN

EXPLOITING MULTIPARALLEL CORPORA AS MEASURE FOR SEMANTIC RELATEDNESS TO SUPPORT LANGUAGE LEARNERS

JOHANNES GRAËN[1] AND GEROLD SCHNEIDER[2]

[1,2] UNIVERSITY OF ZURICH (SWITZERLAND)
[1]UNIVERSITY OF GOTHENBURG (SWEDEN)
[1]POMPEU FABRA UNIVERSITY (SPAIN)

Introduction

Computational Linguistics and Learner Error research have made impressive progress recently, but they have not reached their collaborative potential yet (Granger and Lefer 2016).

Learners could profit from the help of computational tools at any linguistic level, but not all phenomena are equally difficult for learners. In the test data of the grammatical error correction shared task of the CoNLL conference in 2014 (Ng et al. 2014), the most frequent type of error made by learners of English is *wrong collocation or idiom*, which contributes to about 14% of all errors, followed by articles, and prepositions.

We suggest an approach which may help learners to improve their use of idioms and collocations, and linguists and teachers to explore them. Among the error type of collocation and idiom, choosing an unsuitable word due to L1 transfer, including false friends, is a frequent source of error (Dahlmeier and Ng 2011). Our approach can help users to detect, explore and study by example any type of semantic choice.

We present two types of semantic difficulty learners are facing: words that are different in meaning but similar at the surface interlingually (false friends), and intralingually, such as particle verbs in comparison to their base form. It is difficult for learners to know that e.g. German *vorschlagen* (suggest) is opaque, i.e.

semantically very different from German *schlagen* (beat), while German *vorlesen* (read out) is transparent and compositional, as it is semantically very close to German *lesen* (read), and the particle *vor* corresponds to its prepositional meaning.

Idiom and collocation errors often involve choosing a word that is semantically inappropriate, for example because it is orthographically and etymologically similar to a word in the learner's L1. Although in certain contexts they may also be a suitable translation, this type of error is also known as *false friend*. They are a frequent and difficult problem for language learners. In her detailed study of false friends that English learners with Spanish L1 background produce, Varela (2012) uses a list of 100 types of typical false friends, and reports that on average 23.4% of their occurrences are incorrect uses. In total numbers, 579 tokens of the totally 2477 tokens are incorrect uses.

Most resources are in the form of dictionaries (Varela 2011). While they are useful resources, they are on the one hand open and incomplete. On the other hand, not all occurrences of the words in the lists of false friends are incorrect; many of them are only partial false friends. Resources that offer real examples in context are thus useful. Some such resources exist, for example *Linguee* (see Volk et al. 2014 for a comparison), but they require a considerable amount of reading, do not offer nice aggregations or visualizations, and they do not specifically target language learners.

Corpus Material

Our work is based on the FEP9 corpus (Graën 2018: 24 ff), which contains processed texts in 16 languages from a cleaned version of the Europarl corpus. The Europarl corpus (Koehn 2005) in its latest version comprises 15 years of transcripts of the European Parliament sittings. Being designed as training data for statistical machine translation systems, it contains a variety of errors, which we classified and partly corrected, resulting in the CoStEP corpus (Graën, Batinic, and Volk 2014). CoStEP holds approximately 87% of the number of tokens available in Europarl. In addition to correspondences on the level of individual sitting dates available in Europarl, texts in the CoStEP release are grouped by individual speaker contributions.

For the creation of the FEP9 corpus, we extract speaker contributions from the 20+ languages available in CoStEP that are at least available in English, French, German, Italian and Spanish. At most, we use these five and translations into 11 other languages shown in Graën (2018). In addition to the raw tokenized texts, FEP9 holds several layers of annotation, such as part-of-speech tagging and lemmatization, and alignment, namely sentence and word alignment. Word alignment is a technique to identify corresponding tokens in parallel texts (for a detailed description see ibid.: 106 ff). Figure 15-1 illustrates the result of word alignment on a parallel English/Spanish sentence.

Figure 51-1. Word alignment example, Lines connect corresponding tokens

The tool presented in this paper uses 12 of the 16 languages available, namely those where we could reliably assign lemmas to the given word forms.

Methods

Based on lemmatization of the individual languages and word alignment on parallel of each language pair, we derive the so-called lemma distribution matrix described in Graën (2018: 44ff). For every pair of source and target language lemma, this matrix holds the conditional probability p_a of a particular target lemma corresponding to the given source lemma. We do so by calculating the frequency of λ_s being aligned to λ_t (see the numerator) in relation to all cases where λ_t is aligned (in the denominator):

$$p_a(\lambda_t|\lambda_s) = \frac{f_a(\lambda_s,\lambda_t)}{\sum_{\lambda_{t'}} f_a(\lambda_s,\lambda_{t'})} \quad (1)$$

The conditional probability is hence the relative frequency of the source lemma to correspond to the target lemma and, consequently, the sum of all possible target lemmas yields 1 for each source lemma.

The absolute and relative frequencies for the Spanish lemma *vaca* on word alignment between English and Spanish is given in Table 15-1. The most prominent aligned English lemma is *cow* with a share of 82% (i.e. *cow* is the lemma of the English token that is aligned with a Spanish token whose lemma is *vaca* in 82% of the cases that we observe in our corpus). At the very end of the list, where frequencies are very low, lemmatization and alignment errors become more prominent.

Lemma	***f***	***p***
cow	305	82.0
cattle	44	11.8
beef	4	1.1
calf	3	0.8
steer	3	0.8
animal	3	0.8
bull	2	0.5
livestock	2	0.5
Bse	2	0.5
bovine	1	0.3
Cattle	1	0.3
underdone	1	0.3
bullock	1	0.3

Table 15-1. Alignment probabilities for Spanish *vaca* to English

Starting with two lemmas λ_1 and λ_2, we retrieve their respective alignment distribution in the form of conditional lemma probabilities and calculate the overlap in terms of third language lemmas:

$$f_{\cap}(\lambda_1, \lambda_2|\lambda_x) = min\left(f_a(\lambda_1, \lambda_x), f_a(\lambda_2, \lambda_x)\right) \quad (2)$$

$$p_{\cap}(\lambda_1, \lambda_2|\lambda_x) = min\left(p_a(\lambda_x|\lambda_1), p_a(\lambda_x|\lambda_2)\right) \quad (3)$$

$f_{\cap}$ is the lower absolute frequency of these two lemmas being aligned with a third language lemma λ_x. $p_{\cap}$ measures the same property of relative frequencies. The Spanish lemma *vacuno*, for instance, is aligned to *cow* in 9 cases, which makes up 1.5% of its occurrences. $f_{\cap}$ of *vaca* and *vacuno* is thus 9 and $p_{\cap}$ amounts to 1.5%.

In order to measure the entire overlap of both lemma alignment distributions, we calculate the sum of alignment probabilities over all possible third language lemmas. Infrequent lemmas can show distorted probabilities. A third language lemma that occurs only once, but is aligned with both λ_1 and λ_2 at this occurrence, has a probability of 1, for instance. With the aim of downgrading those cases, we use the logarithm of the absolute frequency as additional weight per lemma:

$$O_a(\lambda_1, \lambda_2) = \frac{\sum_{\lambda_x} log(f_\cap(\lambda_1,\lambda_2|\lambda_x)+1)\cdot p_\cap(\lambda_1,\lambda_2|\lambda_x)}{\sum_{\lambda_x} log(f_\cap(\lambda_1,\lambda_2|\lambda_x)+1)+\epsilon} \quad (4)$$

The alignment overlap measure which is thus always between 0 and 1, can be interpreted as a probability, and is defined in a way which captures the semantic similarity of two lemmas given a particular corpus. This is possible since word alignment targets the identification of which tokens of a source language sentence have been translated to which tokens of the target language sentence. In that way "word alignment is able to attach semantic information to word and multiword units, by means of their target language counterparts." (Medeiros Caseli et al. 2010: 61).

Evaluation

We employ a dictionary of false friends to evaluate our method. We rely on two online dictionaries:

- http://mentalfloss.com/article/57195/50-spanish-english-false-friend-words
- https://en.wiktionary.org/wiki/Appendix:False_friends_between_English_and_Spanish

After removing a few trivial and contested cases, our dictionary contains 64 items. False friends are expected to have a low overlap, while the prototypical translations, which we will refer to as *good friends*, are expected to obtain a high overlap score. Examples from the list, with both false and good friends, are given in Table 15-2.

ES	**EN false friend**	**EN Trans = good friend**
actual	actual	current
asistir	assist	attend
campo	camp	countryside
compromiso	compromise	obligation
decepción	deception	disappointment
introducir	introduce	insert
éxito	exit	success
suceso	success	event
recordar	record	remember
vaso	vase	glass

Table 15-2. False friends and good friends of examples from our dictionary

We have set thresholds of 25% and 50%. If each false friend were to lie above the threshold, and each good friend below, our method would achieve full precision. Our results are given in Table 15-3.

Threshold	Precision (false friend)	Precision (good friend)
25%	88.9% (40/45)	70.0% (15/18)
50%	80.7% (46/57)	83.3% (21/30)

Table 15-3. Performance of classification into false friends and good friends, using different thresholds

A threshold of 25% obtains almost 90% precision, while the more balanced threshold of 50% obtains an F-score of above 80%. We explored both precision and recall errors, and found that there are both true errors and cases of partial false friends. We have also developed a graphical interface which allows the user to explore the semantic graph of alignment relations, as we discuss in the following section.

Web Interface For Exploration

We now describe our interactive web interface, which is publically available at http://pub.cl.uzh.ch/purl/alignment_overlap. We first show two examples of German particle verbs. The complex verb *auslösen* (*trigger*) shows no overlap to its base verb without particle *lösen* (*solve*), as we expected as their relation is opaque. The screenshot is given in Figure 15-2. In contrast, the German lemma pair *ansteigen* (*increase*) and *steigen* (*rise*, *climb, increase*) exhibits an almost complete overlap in their translations, as Figure 15-3 shows.

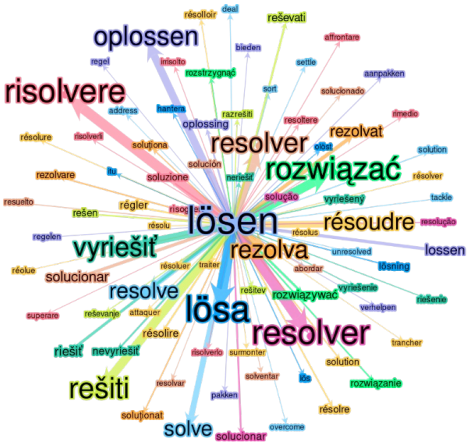

Figure 15-2. German *lösen* and *auslösen* show no overlapping translations

It is worth teaching frequent particle verbs which have no or hardly any overlap separately to broaden learners' vocabulary. Our tool allows teachers and linguists to detect them, and for learners it also reveals the non-compositional meaning of the particle verb.

Figure 15-3. German *steigen* and *ansteigen* overlap very strongly

Let us turn to false friends now. While very strong false friends have no overlap at all, most items in our dictionary show some degree of overlap. *Entender* (understand) is a false friend of French *entendre* (hear). But the separation is not complete, particularly German *verstehen* and English *understand* are sometimes used with the meaning of *hear* if pronunciation is unclear or the sound too low. The partial overlap is shown in Figure 15-4.

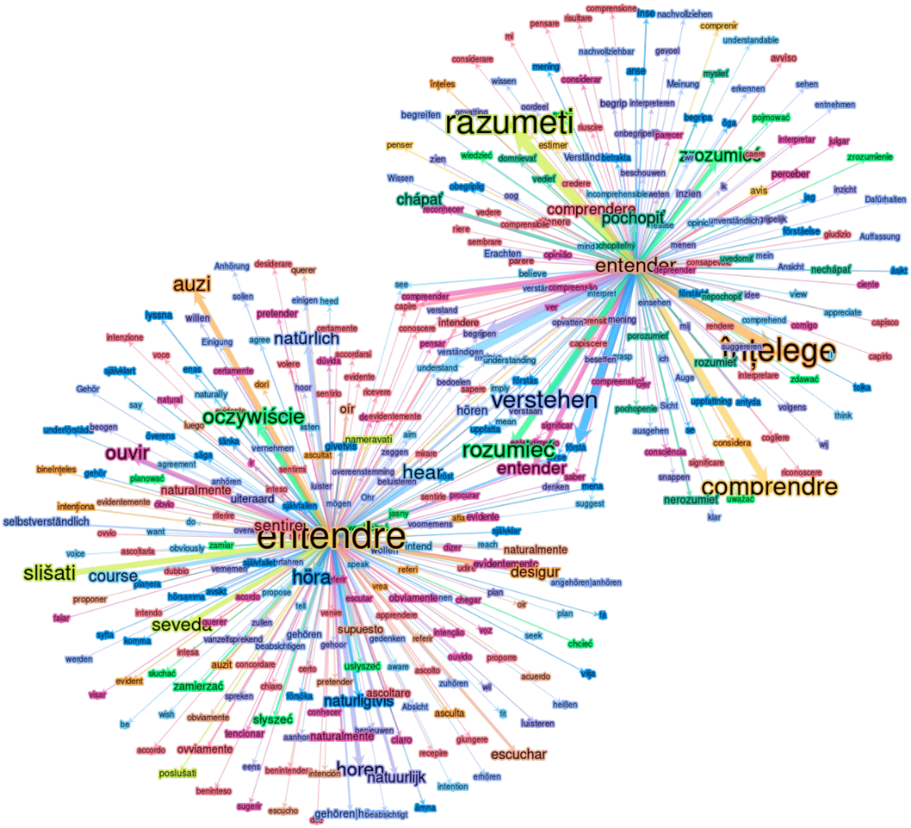

Figure 15-4. French *entendre* and its false friend *entender*

The example of Spanish *tirar* vs. French *tirer* reveals that some instances are shared via the English word *pull*. Looking at the instances, which are retrieved by clicking on the desired arrow in the web interface, shows that most of them are idioms: the English idiom *pull the rug from under one's feet* is translated in Spanish as *tirar de la alfombra*, as Figure 15-5 shows.

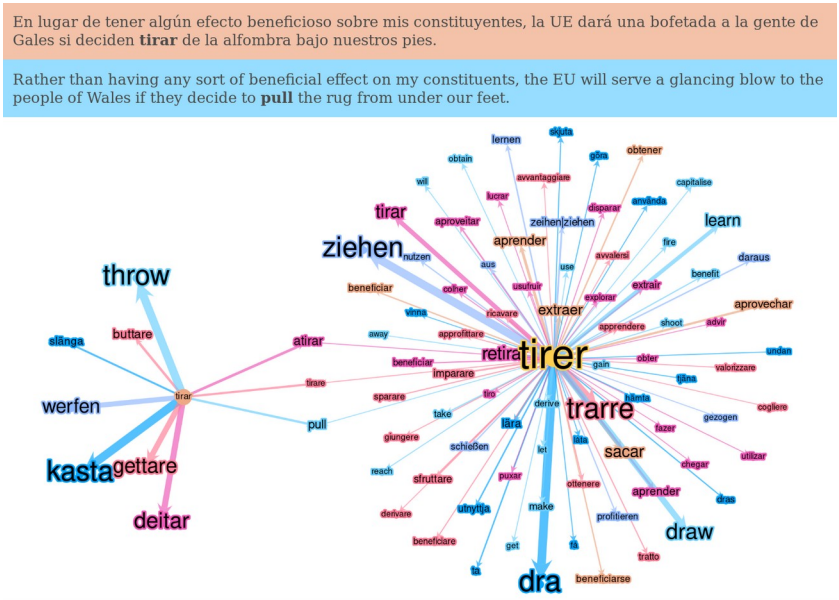

Figure 15-5. Overlap between Spanish *tirar* and its French false friend *tirer*, which is correct in certain idioms

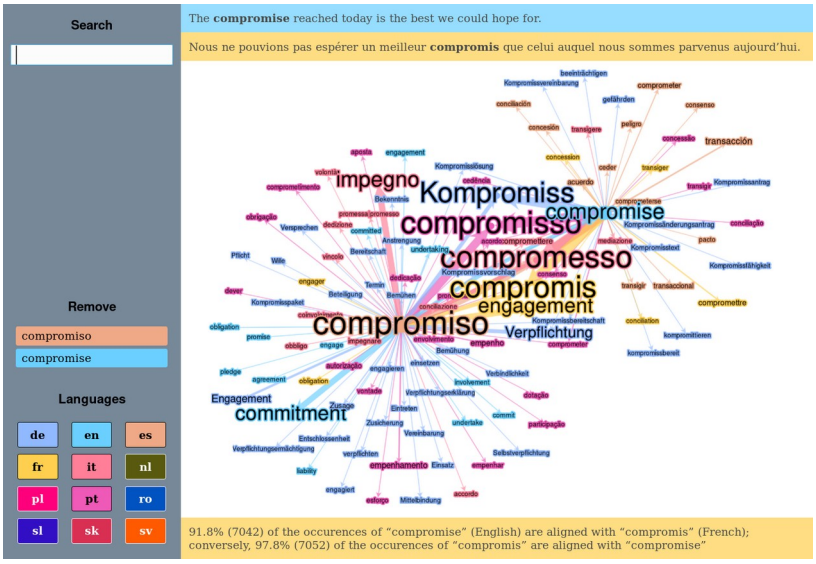

Figure 15-6. Spanish *compromiso* and its English partial false friend *compromise*

Among the false friends that our approach failed to recognize we find *compromiso*, which is claimed to be a false friend of *compromise*, while its correct translation should normally be *obligation*. We see in Figure 15-6 that, via the French *compromise* among others, the alleged false friend translation is used, particularly when a *compromise* in the form of an agreement has been reached, as we can see when browsing the examples (which are shown on the top of the screen when clicking an arrow, see Figure 15-6). We also show the query fields and language selection buttons (ISO language codes). When we add the suggested standard translation of *compromiso* to *obligation* to the picture, we can observe a triangular relationship with connections via third language lemmas between all three terms, as Figure 15-7 illustrates.

Figure 15-7. The triangular relationship between Spanish *compromiso* and the English competing translations *compromise* and *obligation*

As we can see, the relations between several related words and their translations can be explored. Spanish *molestar* does not only have the English false friend *molest*, its translation into English also depends on the exact meaning, as Figure 15-8 shows. It can translate to *annoy*, *disturb* or *bother* with similar likelihood (indicated by arrow thickness), and demands particular attention for translation.

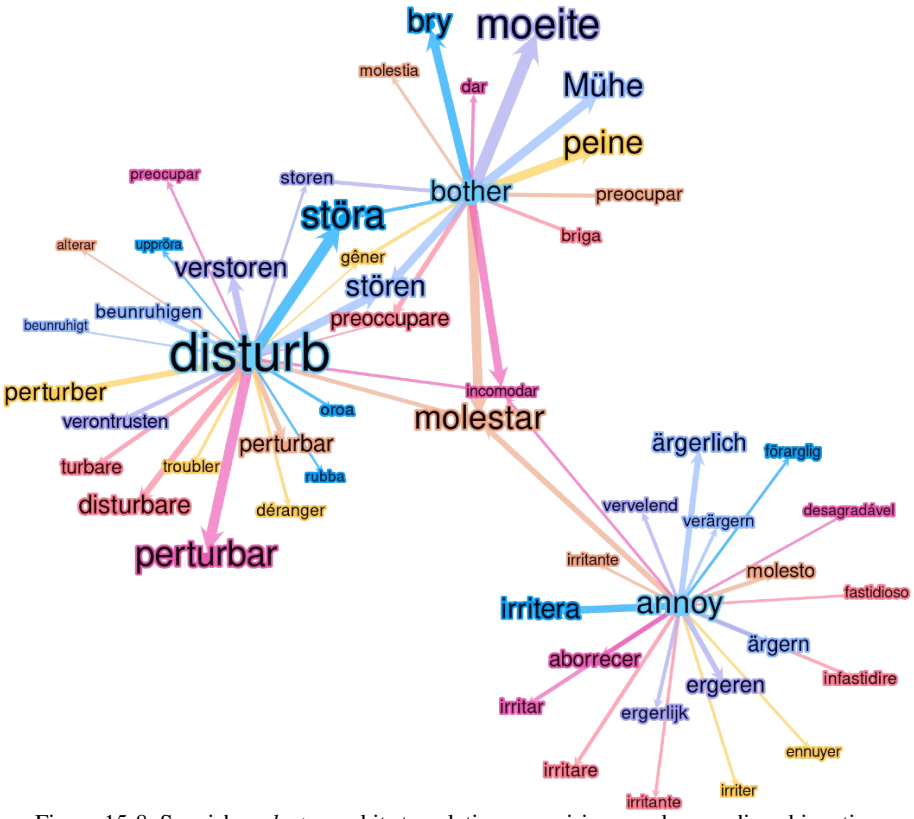

Figure 15-8. Spanish *molestar* and its translations, requiring word sense disambiguation between English *disturb*, *annoy* and *bother*

Conclusions and Future Work

We have presented and evaluated a tool for language learners, teachers and translators, which allows them to find appropriate translations, avoid false friends, explore non-compositional expressions such as particle verbs. Our tool also enables

one to explore translations of semantically related words, for example compositionality of particle verbs, disambiguation via intermediate languages, idiomatic expressions, and more.

Our evaluation against a popular list of false friends delivered a balanced system of above 80% precision and recall, or about 90% at 70% recall. We have discussed cases of partial false friends, such as French *entendre* and Spanish *entender*. Our tool offers the possibility to explore them, strengthening the intuitions of advanced learners.

We plan to evaluate our resource in future research together with translators and language learners in order to find out if they find it useful. We further envisage the following future applications:

- Inclusion of these visualizations in bilingual (or multilingual) web dictionaries, for example *Multilingwis,* which is publicly available at http://pub.cl.uzh.ch/purl/multilingwis
- Automatic analysis of the context of the overlap, e.g. English *course* is aligned to French *entendre* only in the context of "of course" / "bien entendue".
- Based on the context and overlap, point out collocations to the language learner.

References

Dahlmeier, Daniel and Hwee Tou Ng. 2011. "Correcting Semantic Collocation Errors with L1-induced Paraphrases". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP). Edinburgh, United Kingdom: Association for Computational Linguistics, 107–117.

Graën, Johannes. 2018. *Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning*. PhD Thesis. University of Zurich.

Graën, Johannes, Dolores Batinic, and Martin Volk. 2014. "Cleaning the Europarl Corpus for Linguistic Applications". In: Proceedings of the Conference on Natural Language Processing (KONVENS). Stiftung Universität Hildesheim, 222–227.

Granger, Sylviane and Marie-Aude Lefer. 2016. "From general to learners' bilingual dictionaries: Towards a more effective fulfillment of advanced learners' phraseological needs". In: *International Journal of Lexicography*, 279–295.

Koehn, Philipp. 2005. "Europarl: A parallel corpus for statistical machine translation". In: *Machine Translation Summit.* (Phuket). Vol. 5. Asia-Pacific Association for Machine Translation (AAMT), 79–86.

Medeiros Caseli, Helena Medeiros de, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. “Alignment-based extraction of multiword expressions”. In: *Language resources and evaluation* 44.1-2, 59–77.

Ng, Tou Hwee, Mei Siew Wu, Ted Briscoe, Christian Hadiwinoto, Hendy Raymond Susanto, and Christopher Bryant. 2014. “The CoNLL-2014 Shared Task on Grammatical Error Correction”. In: *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL): Shared Task*. Baltimore, Maryland: Association for Computational Linguistics, 1–14.

Varela, Maria Luisa Roca. 2011. “Teaching and Learning “false friends”: a review of some useful tools”. In: *Encuentro* (20), 80–87.

Varela, Maria Luisa Roca. 2012. *New Insights into the Study of English False Friends: Their Use and Understanding by Spanish Learners of English*. PhD Thesis. Universidade de Santiago.

Volk, Martin, Graën, Johannes and Callegaro, Elena. 2014. “Innovations in Parallel Corpus Search Tools”. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, 3172-3178.